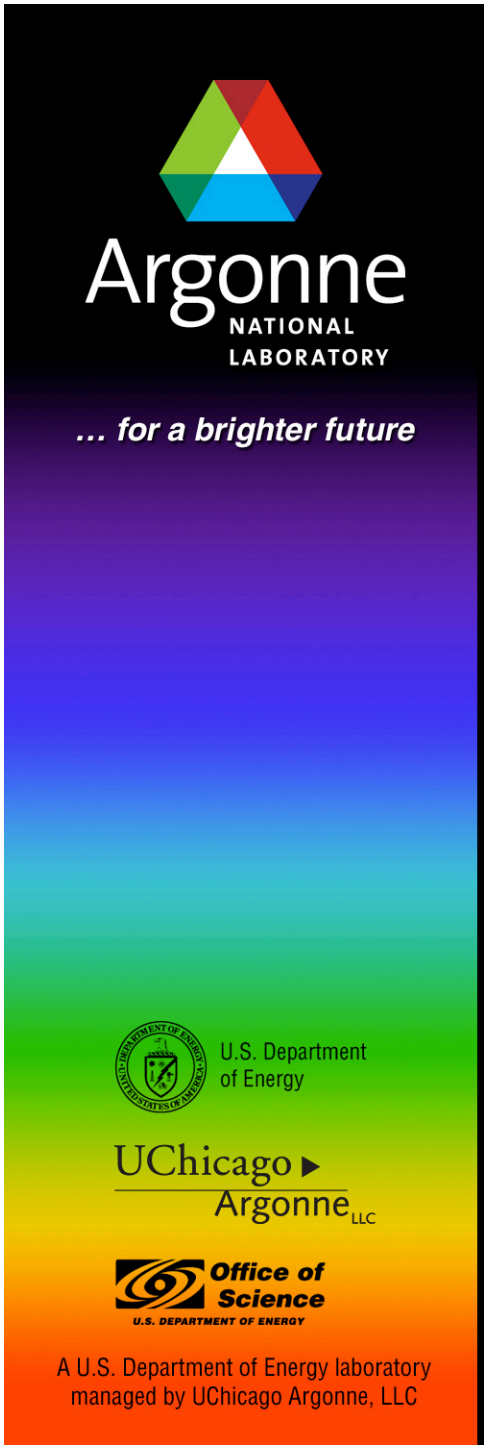


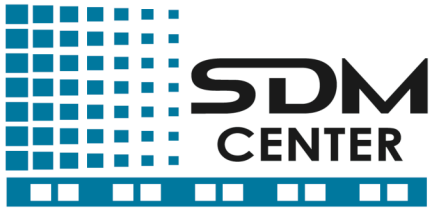
The SciDAC SDM Center:
Moving Research into Production

Rob Ross

Mathematics and Computer Science Division

Argonne National Laboratory

rross@mcs.anl.gov

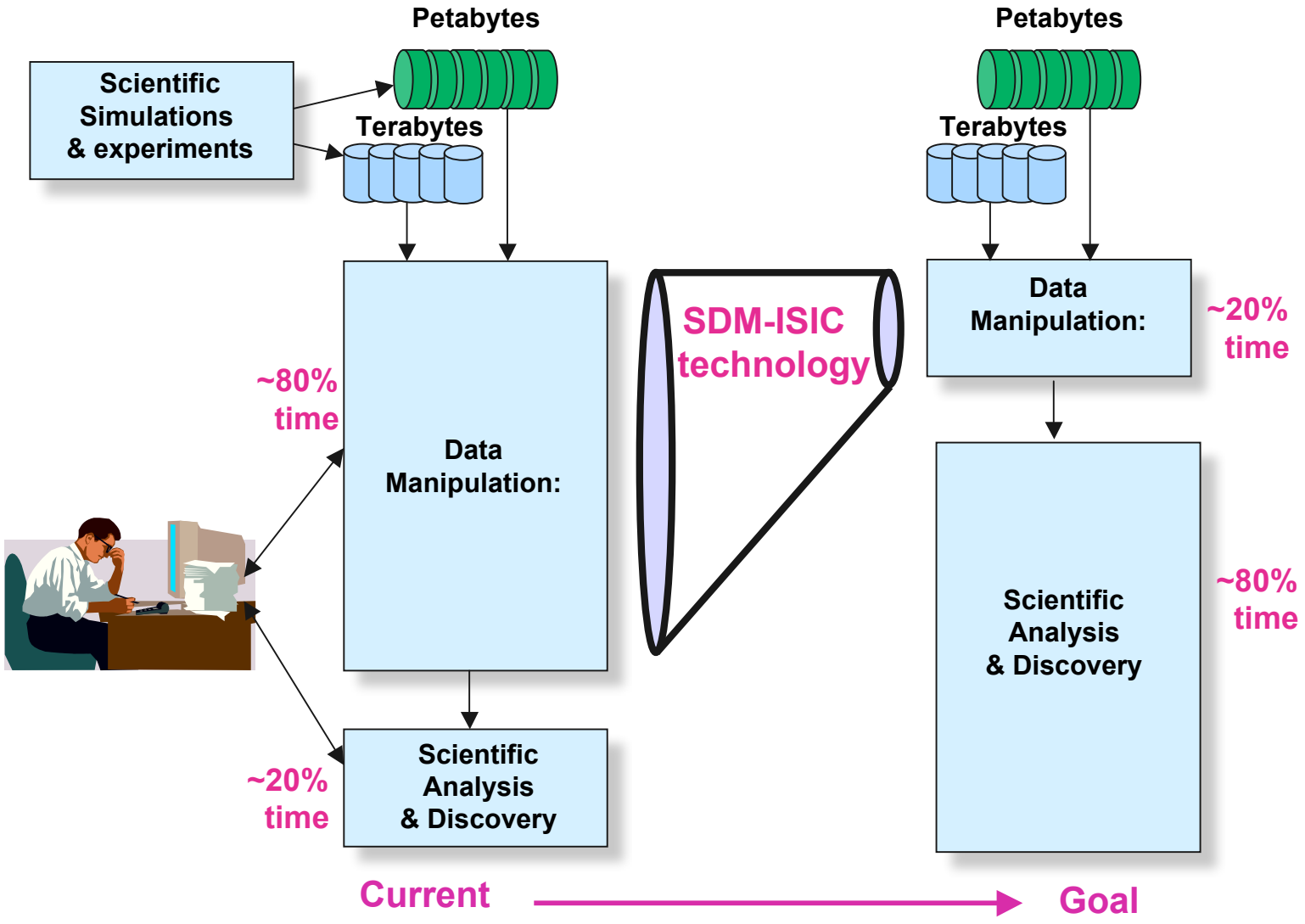


Who is the SDM Center?

- Lawrence Berkeley National Laboratory
  - Arie Shoshani (PI), Doron Rotem
- Argonne National Laboratory
  - Rob Ross, Bill Gropp, Rajeev Thakur
- Lawrence Livermore National Laboratory
  - Chandrika Kamath
- Oak Ridge National Laboratory
  - Jeff Vetter
- Pacific Northwest National Laboratory
  - Jarek Nieplocha, Terence Critchlow
- Northwestern University
  - Alok Choudhary
- North Carolina State University
  - Mladen Vouk, Nagiza Samatova
- University of Utah
  - Steve Parker
- University of California at Davis
  - Bertram Ludaescher
- San Diego Supercomputer Center
  - Ilkay Altinas

- Expertise in a wide range of I/O and data management technologies, from wide-area data movement, to data analytics, to automation of workflows, to I/O middleware and parallel file systems

The SDM Center: Reducing Data Management Overhead



Making Products from Prototypes

- Stages in technology development

  Research ➔ Prototype ➔ Product ➔ Infrastructure

  – Technology starts off as research ideas
  – Technology becomes "infrastructure" when sites start installing it by default
    - *Happens when use in applications hits a certain critical mass*

- One important role of SDM center: Prototype ➔ Product
  – Apply technologies that have been prototyped
  – Fill gaps in I/O software stack on leadership class machines
  – Make application groups aware of these products and their uses

Example: Deploying PVFS at the Argonne Leadership Computing Facility

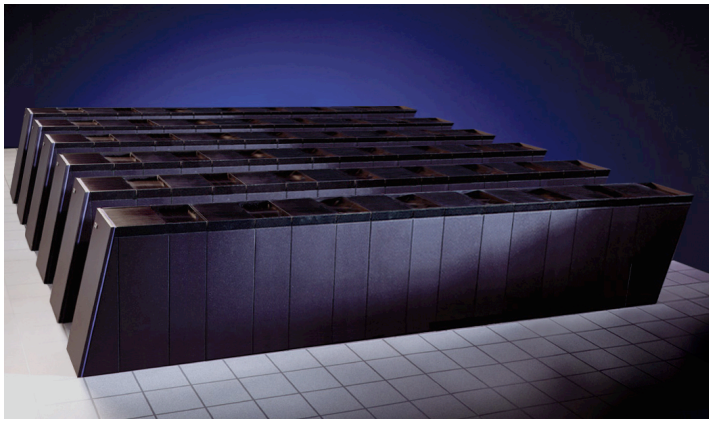
(a work in progress)

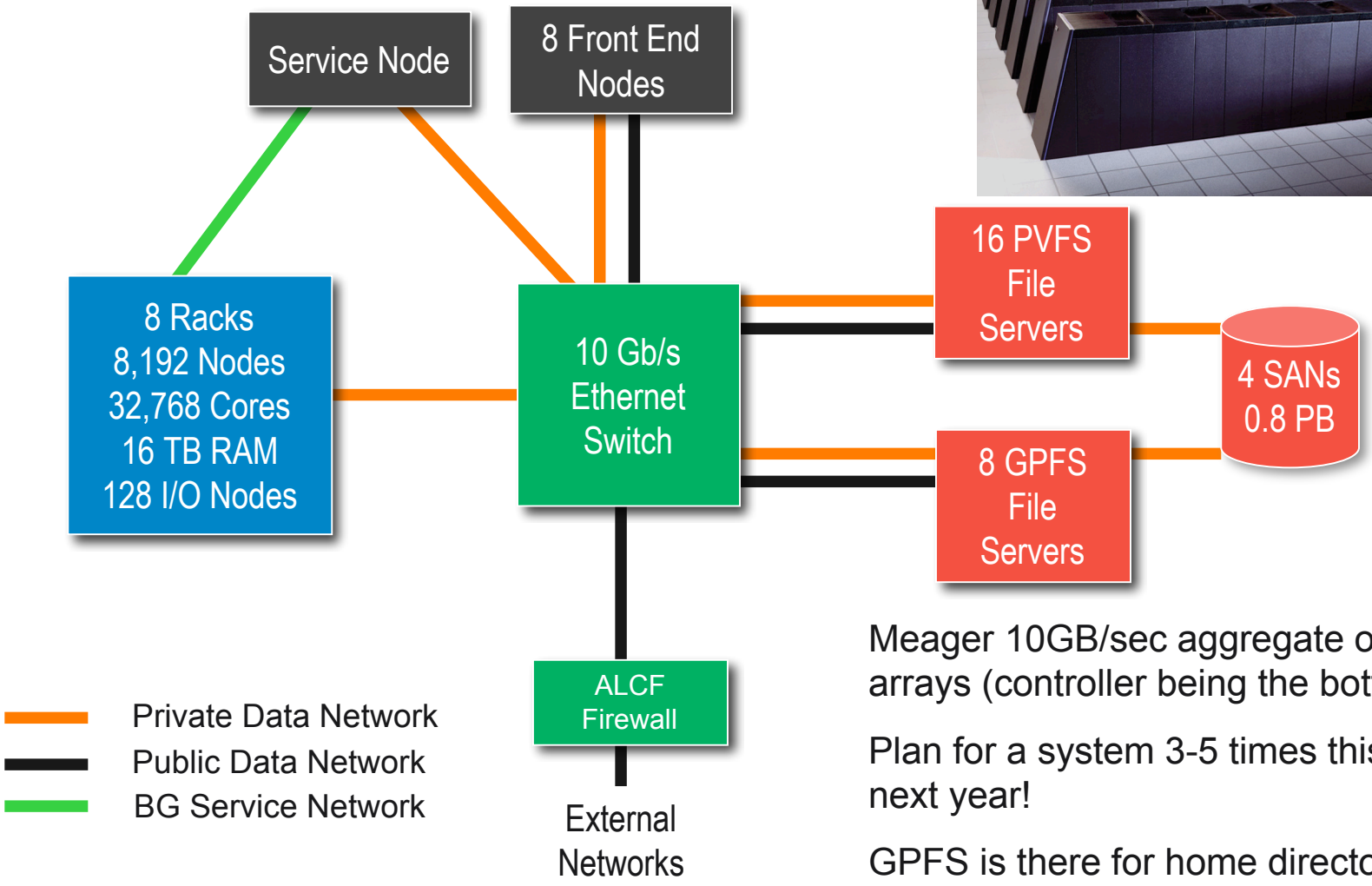
Cheating: We're Already in Production

- "… the world's largest processor of consumer data …" – Fortune magazine
- Multi-national (US, UK, France, Australia, Japan)
- Houses data and runs analytics applications for financial and other large businesses
  - Compute and data intensive
  - 24/7 operation
  - Highly-available, redundant resources
  - 7000+ compute nodes deployed in widely distributed environment
- Deploy PVFS as data storage solution
  - 80 PVFS clusters (16 nodes/cluster)
  - 750TB+ of PVFS storage deployed so far
  - Many internal applications use PVFS libraries directly
- Actively participating in PVFS2 development

ACXIOM®
REDISCOVER YOUR POTENTIAL™

Argonne BG/P "100T" System Architecture





Meager 10GB/sec aggregate out of storage arrays (controller being the bottleneck).

Plan for a system 3-5 times this size early next year!

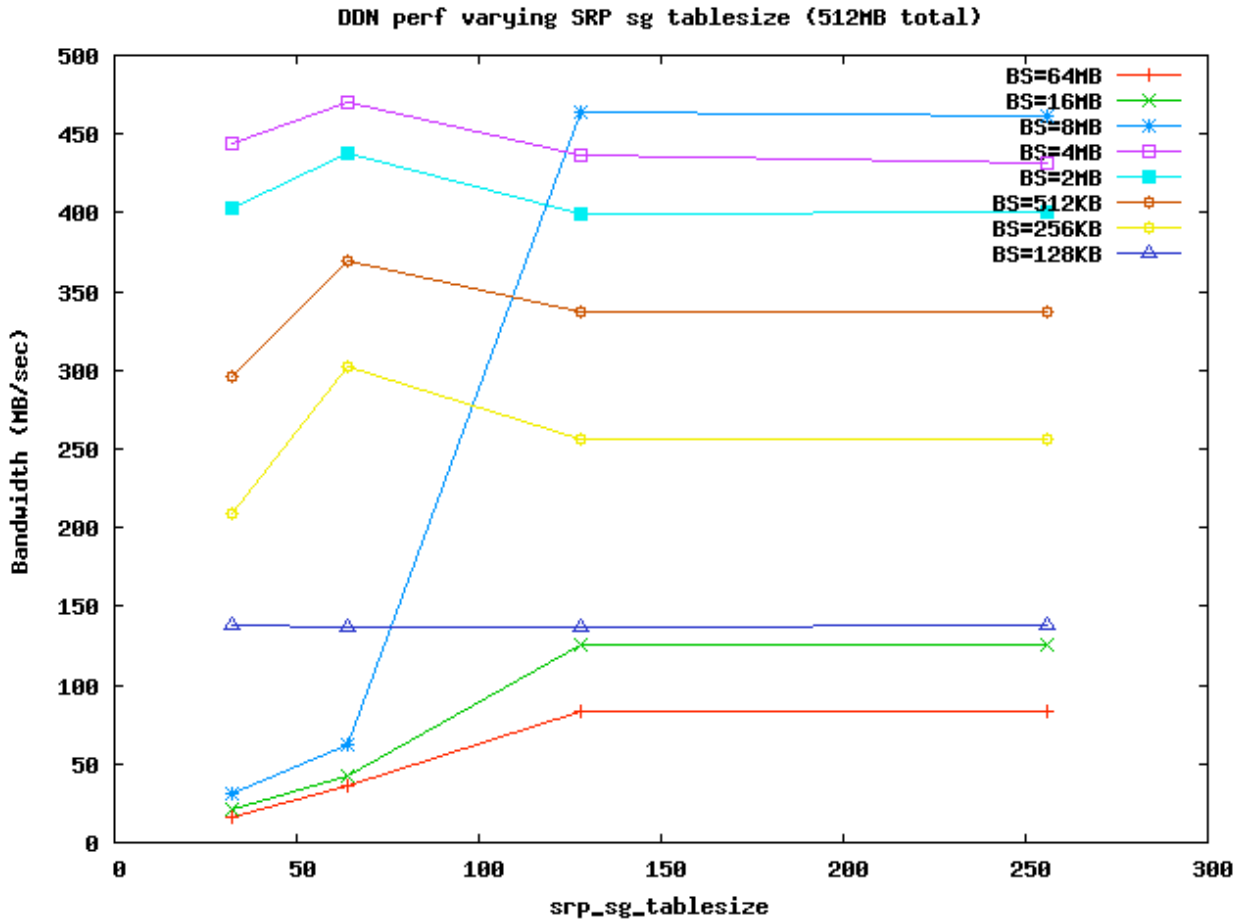
GPFS is there for home directories.

Step 1: Get Educated on Next-Gen Hardware

- As a researcher, I never got to play with the newest, coolest hardware, so eventually I stopped keeping up with it
- We had to quickly educate ourselves on the hardware that would be available in the time frame (storage arrays, InfiniBand rates, PCI-E rates, servers, network gear)
- Needed to be able to estimate bottlenecks and raw hardware rates in order to be able to argue for what was needed
- Worked with ALCF team to specify hardware requirements and begin search for hardware that would meet our needs
- Lots of concalls and NDAs and so forth

Step 2: Tune Our Software on High-End Hardware

- InfiniBand-attached storage does not behave like local disks…

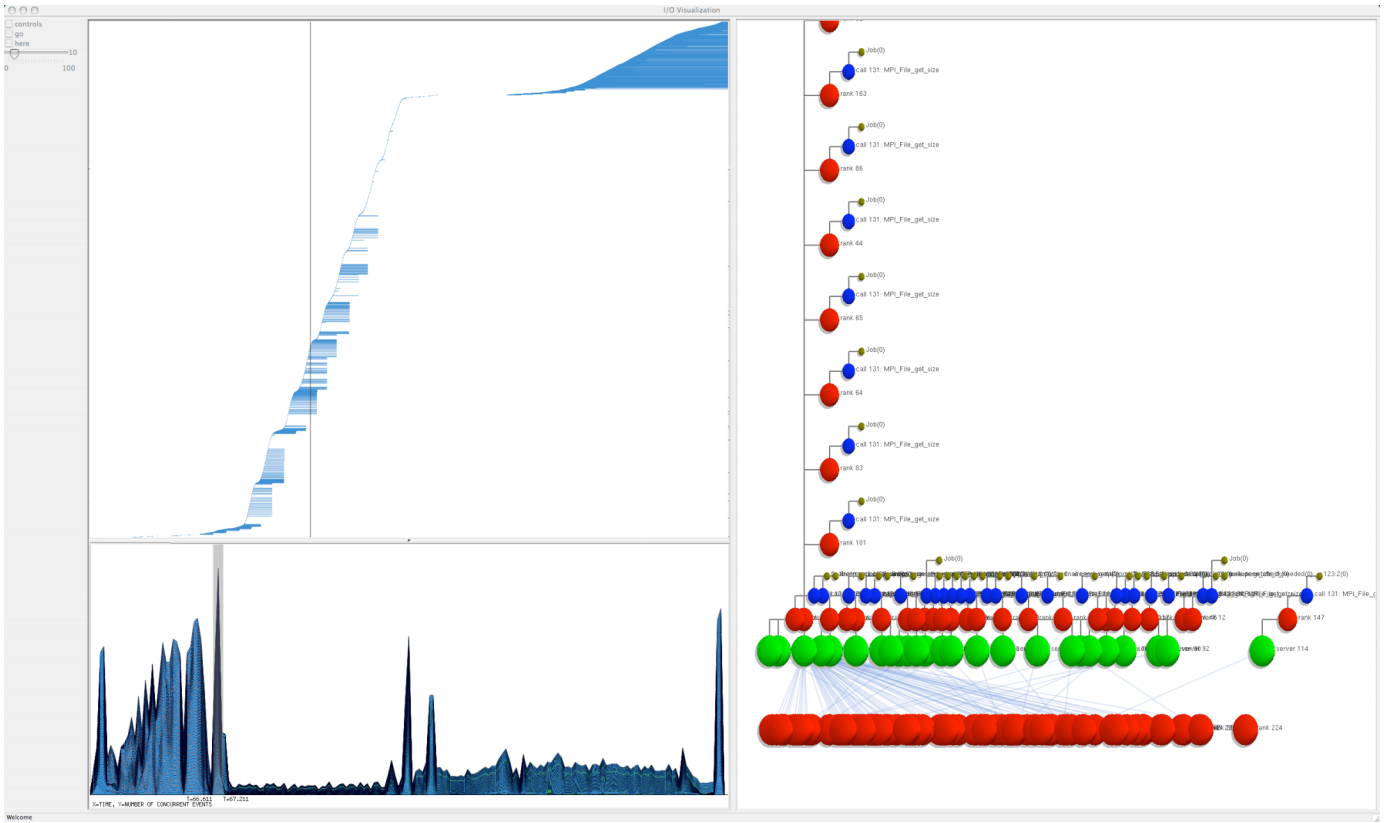


Step 3: Fill in the Gaps (in Progress!)

- Build an efficient MPI-IO for BG/P and PVFS
  - Leverage work at ORNL, NWU, and IBM
- Fix I/O forwarding (ZeptoOS team and others)
- Bypass TCP/IP on the interconnect (with Myricom)
- Get serious about HA (with help from community)
- Avoid incast with a 2D file distribution

Step 4: Prepare for the Worst

- What are the chances that performance will be good out of the box?
- Work with Maloney's TAU team and Ma's InfoVis team to prepare to visualize I/O patterns



How did the SDM Center Help?

- It funds the PVFS side of the work
- It builds collaborations internally, such as with ORNL and NWU
- It put us in contact with the PDSI SciDAC
  - Incast
  - Dirty secrets of storage arrays
- It put us in contact with the TAU team (part of PERI SciDAC) and Ma's InfoVis team (indirectly through the IUSV SciDAC)

Wrapping Up

- Before this experience, I thought I understood how time consuming it could be to support open source software for production use.
- I had no idea. I'm sure that we're in for more surprises.

- On the ALCF side, PVFS expertise is ramping up, and I expect to have a great working relationship with them.
  - Similar to our relationship with Acxiom.
- This help will enable PVFS to be effective infrastructure for ALCF.
- PVFS will be a much stronger tool as a result of this endeavor.

- The experience has been very exciting!
- I can't wait to see it all working!

Many People Make PVFS Possible!

- Collaborators at ANL
  - W. Gropp, R. Thakur, P. Beckman, S. Lang, R. Latham, S. Coghlan, K. Yoshii, and K. Iskra
- Community and industry partners
  - W. Ligon and B. Settlemyer
    Clemson University
  - P. Wyckoff and T. Baer
    Ohio Supercomputer Center
  - P. Carns and D. Metheny
    Acxiom Corporation
  - A. Choudhary and A. Ching
    Northwestern University
  - T. Ludwig and J. Kunkel
    University of Heidelberg
  - D.K. Panda
    Ohio State University
  - P. Honeyman and D. Hildebrand
    University of Michigan
  - L. Ward, R. Klundt, and J. Schutt
    Sandia National Laboratories
  - B. Bode and K. Schochenmaier
    Ames Laboratory